

Biophysical Journal, Volume 118

Supplemental Information

**Distance-Based Metrics for Comparing Conformational Ensembles of
Intrinsically Disordered Proteins**

**Tamas Lazar, Mainak Guharoy, Wim Vranken, Sarah Rauscher, Shoshana J.
Wodak, and Peter Tompa**

Supplementary methods

1. Comparison with variants of the proposed distance metrics

1.1 Evaluating differences between the root mean square averages rather than the medians of the inter-residue distance distributions.

We computed the difference between the root mean square average of the $d(i,j)$ distributions of two ensembles, $Diff_d_{avg}RMS(i,j)$, as follows:

$$Diff_d_{avg}RMS(i,j) = \left| \sqrt{\frac{1}{N} \sum_{i,j} d_A(i,j)^2} - \sqrt{\frac{1}{N} \sum_{i,j} d_B(i,j)^2} \right| \quad (1)$$

where $d(i,j)$ are the distances between residue pairs i,j , N is the total number of conformations in the ensembles, and A and B are the two ensembles that are being compared. The above equation is analogous to Eq (1) of the main text.

This yields the following global measure for the difference between two ensembles, equivalent to the ens_dRMS measure of Eq (4) of the main text:

$$ens_dRMS' = \sqrt{1/n \sum_{i,j} Diff_d_{avg}RMS(i,j)^2} \quad (2)$$

with n representing the number of i,j residue pairs.

Heatmaps obtained for pairs of experimentally derived IDP/IDR ensembles, using respectively, $Diff_d_{avg}RMS(i,j)$ and $Diff_d\mu(i,j)$, of Eq (1) of the main text, displayed virtually identical patterns, as illustrated for the $E1$ and $E2$ ensembles of tau-K18 (**Supplementary Figure S1A,B**). For these ensembles the correlation between $d_{avg}(i,j)$ and $d\mu(i,j)$ values was very high (Pearson's $r > 0.99$). The main difference was that the heatmap features computed using $Diff_d\mu(i,j)$ showed somewhat better contrast than those computed with $Diff_d_{avg}RMS(i,j)$. This may be explained by the fact that the root mean square average values tend to be

affected by a few outlier values, whereas the median values are not. The latter are more robust since they represent the most populated $d(i,j)$ value.

A high correlation (Pearson's $r = 0.89$ and 0.87) was also obtained between the global measures e.g. those of ens_dRMS vs ens_dRMS' , computed for the 10 pairs between 5 experimentally characterized human tau-K18 ensembles.

Virtually the same results were obtained when we simply computed the difference between the average values of the $d(i,j)$ distributions of two ensembles instead of Eq (1) above :

$$Diff_d_{avg}(i,j) = \left| \frac{1}{N} \sum_{i,j} d_A(i,j) - \frac{1}{N} \sum_{i,j} d_B(i,j) \right| \quad (3)$$

where A and B are the two ensembles and N is the number of conformations in the ensembles.

This analysis confirms that our approach could readily accommodate measures based on the average values of the $d(i,j)$ distributions, with negligible effects on the results.

1.2 Analyzing differences between $C\beta$ - $C\beta$ distance distributions instead of those between $C\alpha$ - $C\alpha$ atoms.

See main text for details and results illustrated in **Supplementary Figure S2A,B**.

2. Comparisons to other distance dependent metrics

2.1 Comparing differences of ensemble averaged R_g with ens_dRMS values for IDP ensembles of our dataset.

The radius of gyration, R_g , of individual conformers within ensembles is computed as described in the Methods section (main text). The difference between average R_g values of two ensembles is computed as:

$$Diff_ensRg = |<Rg(A)> - <Rg(B)>| \quad (4)$$

where $< >$ indicates averages over conformations in an ensemble, and A and B are different ensembles.

Using the 5 experimentally characterized IDP/IDR ensembles of respectively, the tau-K18 and MeV N-tail proteins of our dataset, representing 10 pairwise comparisons for each system, we computed the $Diff_ensRg$ and ens_dRMS quantities for all 10 ensemble pairs, with results listed in **Supplementary Table S1**. Scatter plots of the $Diff_ensRg$ versus ens_dRMS values for these ensemble pairs are shown in **Supplementary Figure S3A**.

2.2 Comparing differences of distance dependent R_{struct} values to ens_dRMS for IDP ensembles of our dataset.

Following Kuzmanic et al. [1] the distance-dependent pairwise RMS value between two conformations/structures K and L was computed as follows

$$dRMS(K, L) = \sqrt{\frac{1}{N} \sum_{i,j} (d_K(i, j) - d_L(i, j))^2} \quad (5)$$

where $d_K(i, j)$ and $d_L(i, j)$ are inter-residue distances of equivalent residues pairs in conformations K and L, and N is the total number of distances.

The distance-dependent ensemble $dRMS$ is computed as follows:

$$\sqrt{\langle dRMS^2 \rangle} = \sqrt{\frac{1}{M} \sum_{K,L} dRMS(K, L)^2} \quad (6)$$

where K and L are pairs of conformations and M is the number of such pairs.

The distance-dependent structural radius of the ensemble is computed as:

$$dR_{struct} = \frac{1}{\sqrt{2}} \sqrt{\langle dRMS^2 \rangle} \quad (7)$$

and the quantity $Diff_dR_{struct}$ is computed as the difference between the dR_{struct} values of the two ensembles, A and B that are being compared:

$$Diff_dR_{struct} = |dR_{struct}(A) - dR_{struct}(B)| \quad (8)$$

Using the same dataset as in Section 1.2, we computed the $Diff_dR_{struct}$ and ens_dRMS quantities for all 10 ensemble pairs, with results listed in **Supplementary Table S2A,B**. The scatter plots of $Diff_dR_{struct}$ versus ens_dRMS for these ensemble pairs are shown in **Supplementary Figure S3B**. The scatter plots of dR_{struct} versus $Diff_ensRg$ values for individual ensembles are depicted in **Supplementary Figure S3C**.

Supplementary Table S1: Comparison of $Diff_ensRg$, $Diff_dR_{struct}$ and ens_dRMS values computed for pairs of experimentally characterized IDP ensembles of respectively, the tau-K18 and MeV N-tail protein segments.

Ensembles	Diff_ensRg	Diff_dR_{struct}	ens_dRMS
tau_E1-E2	0.93	0.94	1.91
tau_E1-E3	0.05	0.08	1.72
tau_E1-E4	0.1	0.2	1.83
tau_E1-E5	1.33	0.73	1.98
tau_E2-E3	0.98	0.86	2.15
tau_E2-E4	0.83	0.74	1.84
tau_E2-E5	0.4	0.21	1.93
tau_E3-E4	0.15	0.12	1.47
tau_E3-E5	1.38	0.65	2.06
tau_E4-E5	1.23	0.53	2.04

N-tail_E1-E2	1.23	0.13	2.83
N-tail_E1-E3	1.16	0.5	2.9
N-tail_E1-E4	1.02	0.44	2.43
N-tail_E1-E5	0.31	0.82	1.62
N-tail_E2-E3	0.07	0.37	1.74
N-tail_E2-E4	0.21	0.31	1.48
N-tail_E2-E5	0.92	0.69	2.1
N-tail_E3-E4	0.14	0.06	1.82
N-tail_E3-E5	0.85	0.32	2.14
N-tail_E4-E5	0.71	0.38	1.75

The Pearson correlations between the 20 values of the 3 different measures in **Supplementary Table S1** are: $Diff_ensRg/ens_dRMS$ ($r=0.68$); $Diff_ensRg/Diff_dR_{struct}$ ($r=0.53$); $Diff_dR_{struct}/ens_dRMS$ ($r=0.07$). The low correlation for the latter two values is due to the poor correlation for values computed for the N-tail ensembles ($r=-0.14$). A significantly higher correlation is obtained for the tau-K18 ensembles ($r=0.66$). The poor correlation for the N-tail ensembles stems from the outlier behaviour of the E1 N-tail ensemble, which features the lowest $\langle Rg \rangle$ value, but near average dR_{struct} value (**Supplementary Figure S3D**). By removing the 4 data points corresponding to N-tail E1, the correlation between $Diff_dR_{struct}$ and ens_dRMS increases to $r=0.56$.

2.3 Comparison with metrics based on the Kullback–Leibler divergence (KLD) of two distributions

To compute the Kullback–Leibler divergence (KLD) of the $d(i,j)$ distributions in our dataset of experimentally derived IDP ensembles we used the KLD formulation for normal distributions [2]. This is an approximation, given that only ~65% of the $d(i,j)$ values are normally distributed. To quantify the difference between $d(i,j)$ distributions in ensembles A and B, we computed the symmetrized form of the KLD distance distributions, $KLD_d(i,j)$ as follows:

$$\text{symKLD}_d(i,j) = (\text{KLD}_d(i,j)(A \parallel B) + \text{KLD}_d(i,j)(B \parallel A)) / 2 \quad (9)$$

The root-mean-square $\text{symKLD}_d(i,j)$ differences between 2 ensembles, the ensKLD , was computed as follows:

$$\text{ensKLD} = 1/n \sum_{i,j} \text{symKLD}_d(i,j) \quad (10)$$

where i,j are individual residue pairs, and n is the number of such pairs.

Results obtained using these formulations, and applying no corrections for small sample size ($d(i,j)$ distance distributions for the experimentally determined IDP ensembles of our dataset comprise only ~200 data points, representing the number of conformations in individual experimentally restrained ensembles), are illustrated in **Supplementary Figure S4 and Supplementary Table S2**.

Moderate Pearson correlation coefficients ($r = 0.42, 0.51$) were observed between the $\text{Diff}_d\mu(i,j)$ and the symmetrized KLD values (Eq (9)) for individual $d(i,j)$ distributions of the tau-K18 ensemble pairs (such as E1/E2, and E2/E3) exhibiting significantly different $d(i,j)$ distributions, as evaluated by the Mann–Whitney–Wilcoxon test (**See main text and Supplementary Figure S4A,B**). But a negligible correlation was observed between the two values for the ensemble pair E3/E4 with no significantly different $d(i,j)$ distributions (see **Supplementary Figure S4C**). A rather high correlation (Pearson's $r = 0.81$) was obtained between the ens_dRMS , and ensKLD (the ensemble averaged $\text{symKLD}_d(i,j)$ values of (Eq (10)) across all 10 pairs of tau-K18 ensembles (**Supplementary Table S2**).

Supplementary Table S2: Comparison of ensKLD and ens_dRMS values computed for pairs of experimentally characterized ensembles of the tau-K18 disordered protein segment.

Ensembles	<i>ensKLD</i>	<i>ens_dRMS</i>
Tau_E1-E2	0.0089	1.91
Tau_E1-E3	0.0063	1.72
Tau_E1-E4	0.0060	1.83
Tau_E1-E5	0.0079	1.98
Tau_E2-E3	0.0099	2.15
Tau_E2-E4	0.0077	1.84
Tau_E2-E5	0.0054	1.93
Tau_E3-E4	0.0042	1.47
Tau_E3-E5	0.0081	2.06
Tau_E4-E5	0.0078	2.04

References

1-Kuzmanic, A., and B. Zagrovic. 2010. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal* 98:861-871.

2- Roberts SJ & Penny W. (2002) Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing* 50(9):2245 - 2257

3-Ozenne, V., R. Schneider, M. Yao, J. R. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge. 2012. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *Journal of the American Chemical Society* 134:15138-15148.

4-Ozenne, V., F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, S. Segard, P. Bernado, C. Charavay, and M. Blackledge. 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28:1463-1470.

Supplementary figure captions

Figure S1:

Comparing $Diff_d_{avg}RMS(i,j)$ to $Diff_d\mu(i,j)$ matrices.

(A) Heat plots depicting difference matrices for the tau-K18 E1 and E2 ensembles computed using $Diff_d\mu(i,j)$, based on the medians of the $d(i,j)$ distributions (upper triangle), and using $Diff_d_{avg}RMS(i,j)$, based on the root-mean-square average of the $d(i,j)$ distributions (lower triangle)

(B) Heat maps highlighting only the statistically significant portions of the two matrices (see Methods section of the main text for details).

Figure S2:

$Diff_d\mu(i,j)$ matrices computed using $C\alpha$ - $C\alpha$ and $C\beta$ - $C\beta$ distance distributions.

(A) Heatmaps depicting difference matrices computed using $C\alpha$ - $C\alpha$ distance distributions. Upper triangle: $\%Diff_d\mu(i,j)$ values, representing normalized differences of the $d(i,j)$ distribution means; lower triangle: $Diff_d\sigma(i,j)$ values, representing differences in standard deviations of the corresponding distributions (see Methods section of the main text for details)

(B) The same plots as in (A), but with $\%Diff_d\mu(i,j)$ and $Diff_d\sigma(i,j)$ values computed using $C\beta$ - $C\beta$ distance distributions.

Figure S3:

Scatter plots illustrating the correlations between the $Diff_ensRg$, $Diff_dR_{struct}$ and ens_dRMS , quantities.

(A) Scatter plot of $Diff_ensRg$ versus ens_dRMS values computed for the 10 pairs of the N-Tail, and tau-K18 ensembles.

(B) Scatter plot of $Diff_dR_{struct}$ versus ens_dRMS values computed for the same ensemble pairs as in (A).

(C) Scatter plot of $Diff_ensRg$ versus $Diff_dR_{struct}$ values computed for the same ensembles as in (A) and (B).

(D) Scatter plot of ensemble-averaged Rg values, $\langle Rg \rangle$, versus dR_{struct} values computed for the 5 tau-K18 ensembles. The E1 outlier ensemble is highlighted with a red circle.

The Pearson correlation coefficient computed between pairs of values in each plot are listed below the corresponding plot.

Figure S4:

Comparison of $symKLD_d(i,j)$ versus $Diff_d\mu(i,j)$ matrices

Heatmaps illustrating examples of difference matrices computed for 3 pairs of tau-K18 of ensembles. Shown are matrices for the E1/E2 (A) and E2/E3 (B) pairs, with statistically significant differences $d(i,j)$ distributions, and for the E3/E4 pair (C), where most of the $d(i,j)$ distributions are not significantly different. The upper triangle of the heatmaps/matrices display $Diff_d\mu(i,j)$ values (Å), and the lower triangles depict values of $KLD'_d(i,j) = (symKLD_d(i,j) \times 100)^2$. The latter quantity was used to increase contrast, allowing the two matrices to be depicted simultaneously using a common color scale.

The Pearson correlation between $symKLD_d(i,j)$ and $Diff_d\mu(i,j)$ values for each pair of ensembles is listed at the bottom of the corresponding heatmap.

Figure S5:

Amino acid sequences of the measles virus (MeV) N-tail and tau-K18 segments, whose conformational ensembles (E1-E5) were analyzed in this study.

Figure S6:

Heat maps highlighting the $Sig_Diff_d\mu(i,j)$ values for the 10 pairwise combinations of the 5 human tau-K18 ensembles (E1-E5). These values represent elements of the $Diff_d\mu(i,j)$ matrices corresponding to statistically

significant differences between the corresponding $d(i,j)$ distributions ($p < 0.05$) (upper triangle), and the corresponding $Diff_d\sigma(i,j)$ values (lower triangle).

Figure S7:

Comparisons of experimentally characterized MeV N-tail IDR ensembles.

Quantifying the similarity between the E2/E4 and E1/E3 pairs of MeV N-tail ensembles displaying, respectively, the smallest (1.48 Å) and largest (2.90 Å) *ens_dRMS* value in **Table 1**. These ensembles were generated as described in references [3,4] of the main text, by creating a very large number of random coil conformations, followed by selection of a subset of conformations (here 199 conformations) that optimized the fit to NMR data.

Panel I: Results for the E2/E4 pair. Top: heat maps of $d\mu(i,j)/d\sigma(i,j)$ matrices for the individual *E2* and *E4* ensembles. Middle left: heat maps of the $Diff_d\mu(i,j)/Diff_d\sigma(i,j)$ computed for the *E2/E4* pairs, featuring several small regions with differences $> 4.85\text{Å}$; middle right: heat maps depicting only the statistically significant elements of these maps ($Sig_Diff_d\mu(i,j)/Sig_Diff_d\sigma(i,j)$), and showing none of the $Diff_d\mu(i,j)$ elements to be statistically significant. Bottom: histogram of the distributions of the gyration radii (R_g) of *E2* and *E4*, found to be statistically indistinguishable ($p=0.3$).

Panel II Results for E1/E3 pair. The top, middle, and bottom panels display the same quantities as in Panel I, computed for this most different pair. The $Diff_d\mu(i,j)$ and $Diff_d\sigma(i,j)$ matrices computed for this pair feature much more prominent difference than those of the E2/E4 pair. The statistically significant elements ($Sig_Diff_d\mu(i,j)/Sig_Diff_d\sigma(i,j)$) highlight significant differences in the distance distributions between a short N-terminal segment and a longer C-terminal region. The R_g distributions of E1/E3 pair (bottom plot) are likewise statistically indistinguishable ($p=0.291$).

Figure S8:

Heat maps highlighting the $Sig_Diff_d\mu(i,j)$ values for the 10 pairwise combinations of the 5 MeV N-tail IDR N-tail ensembles (E1-E5). These values represent elements of the $Diff_d\mu(i,j)$ matrices corresponding to statistically significant differences between the corresponding $d(i,j)$ distributions ($p < 0.05$) (upper triangle), and the corresponding $Diff_d\sigma(i,j)$ values (lower triangle).

Figure S9:

Distributions of the radius of gyration for the ensembles of the intrinsically disordered (SR)-rich peptide generated using MD simulations with 5 different force fields (see Methods for detail).

(A) AMBER (03w)

(B) C22: CHARMM22*

(C) Absinth: CAMPARI using the ABSINTH implicit solvent model

(D) AMBER (99sb*-ildn)

(E) C36: CHARMM36

(F) High-T (high temperature)

Figure S10:

Secondary structure classification based on the subdivision of the Ramachandran map adapted from Ozenne et al. 2012, J. Am. Chem. Soc. [DOI: 10.1021/ja306905s].

α : α -helix; β : β -strand; PPI: poly-proline I/II; LH: left-handed helix.

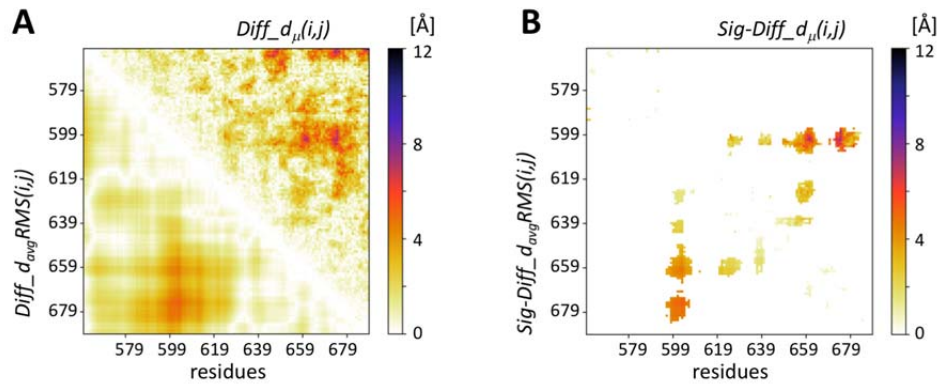


Figure S1

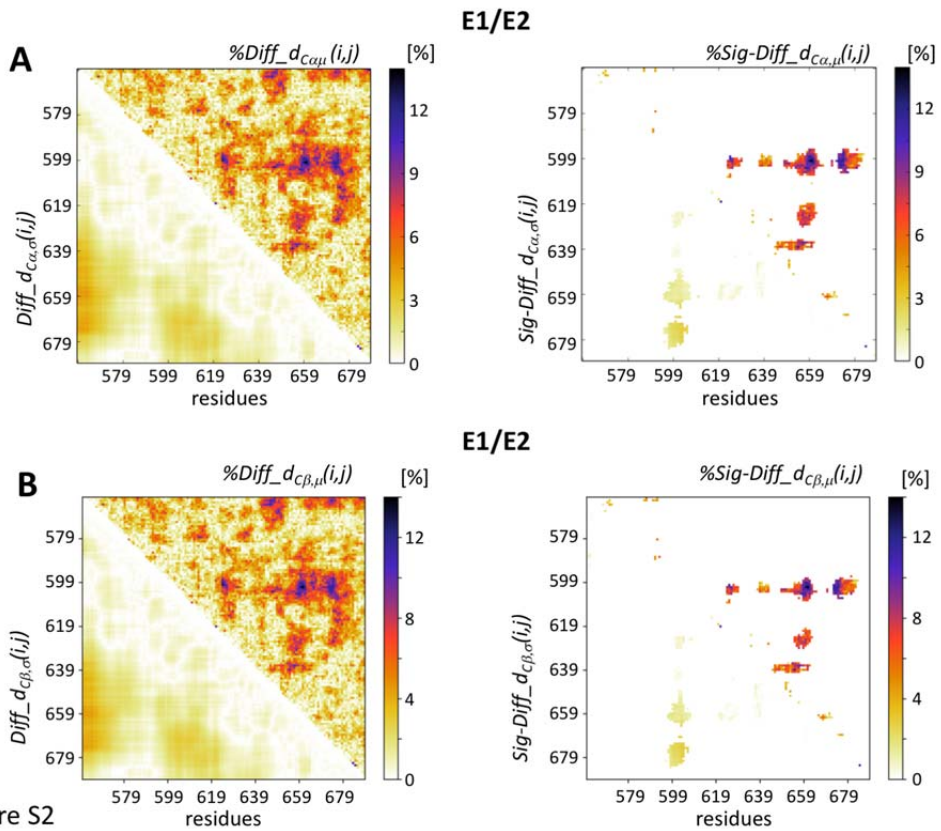


Figure S2

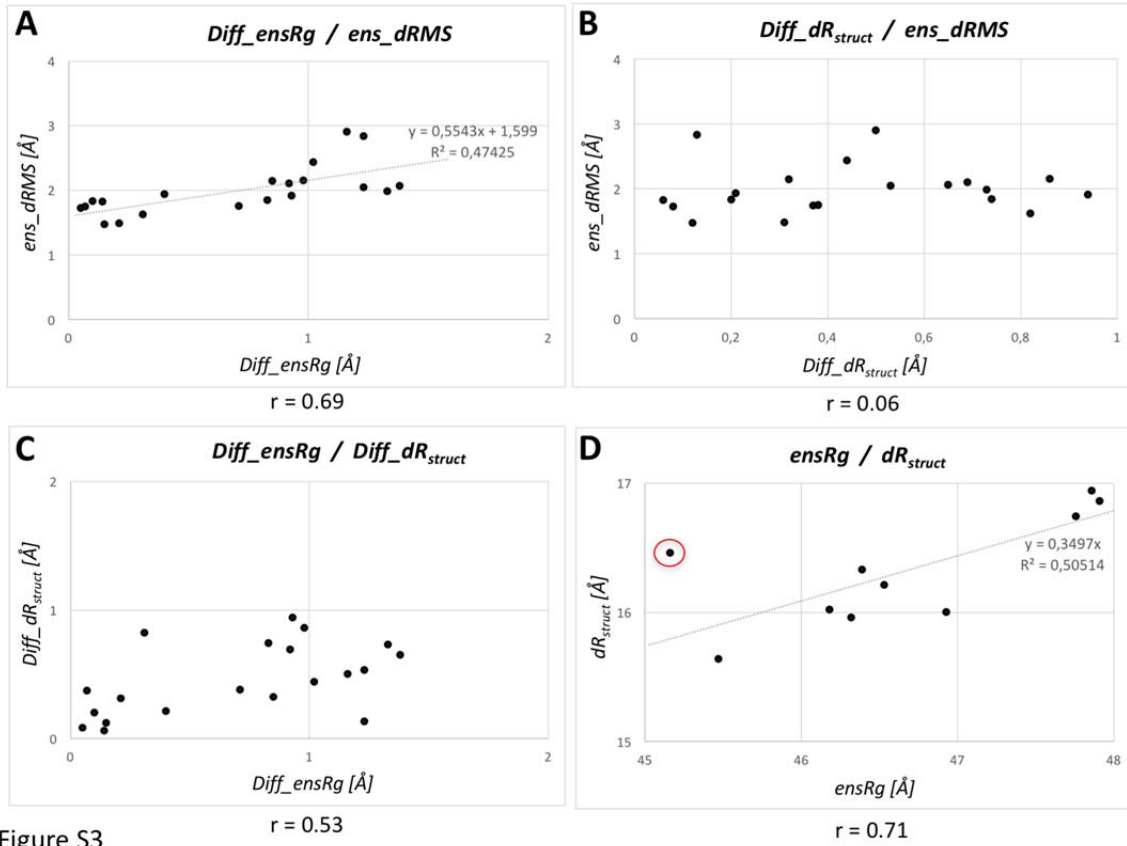


Figure S3

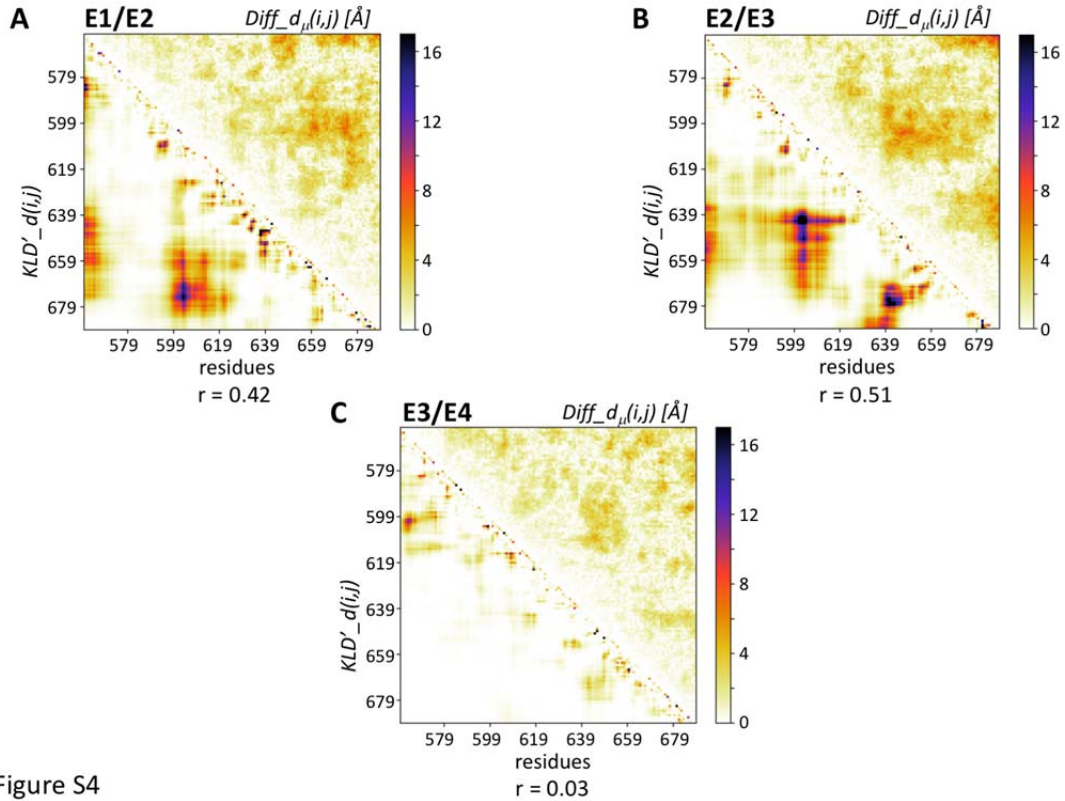


Figure S4

AA sequences of the MeV N-tail and tau-K18 domain

```
>PED7AAC|NCAP_MEASF Nucleoprotein N-tail OS=Measles virus
MHHHHHTTEDKISRAGPRQAQVSFLHGDQSENELPRLGGKEDRRVKQSRGEARESYRET
GPSRASDARAHLPTGTPLDIDTASESSQDPQDSRRSADALLRLQAMAGISEEQGSDTDP
IVYNDRNLLD
```

```
>PED6AAC|TAU_HUMAN MAPT tau-K18 segment OS=Homo sapiens
LQTAPVPMPLKKNVSKIGSTENLKHQPGGGKVQI INKKLDLSNVQSKCGSKDNIKHVPGG
GSVQIVYKPVDLKSVTSKCGSLGNIHHPGGGQVEVKSEKLDKDRVQSKIGSLDNITHVP
GGNKKIE
```

Figure S5

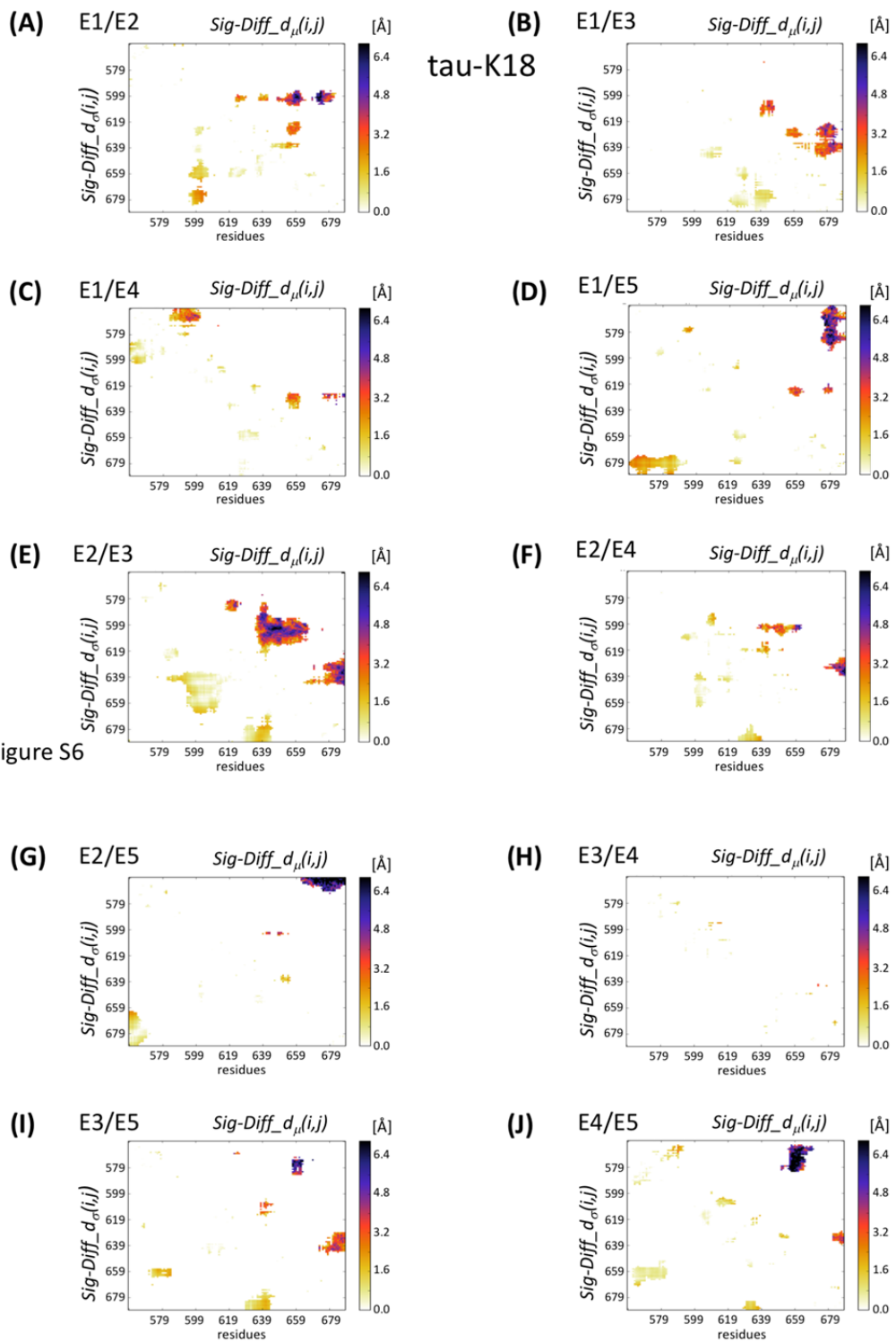


Figure S6

Figure S6

N-tail

[I]

[II]

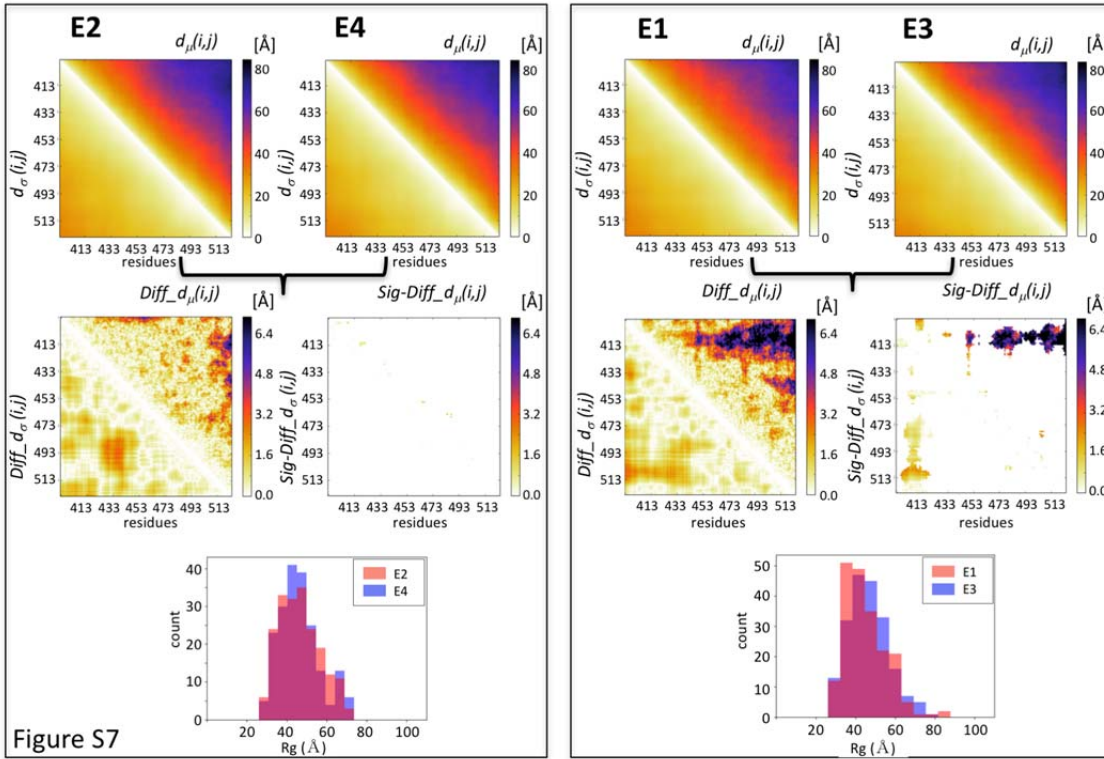


Figure S7

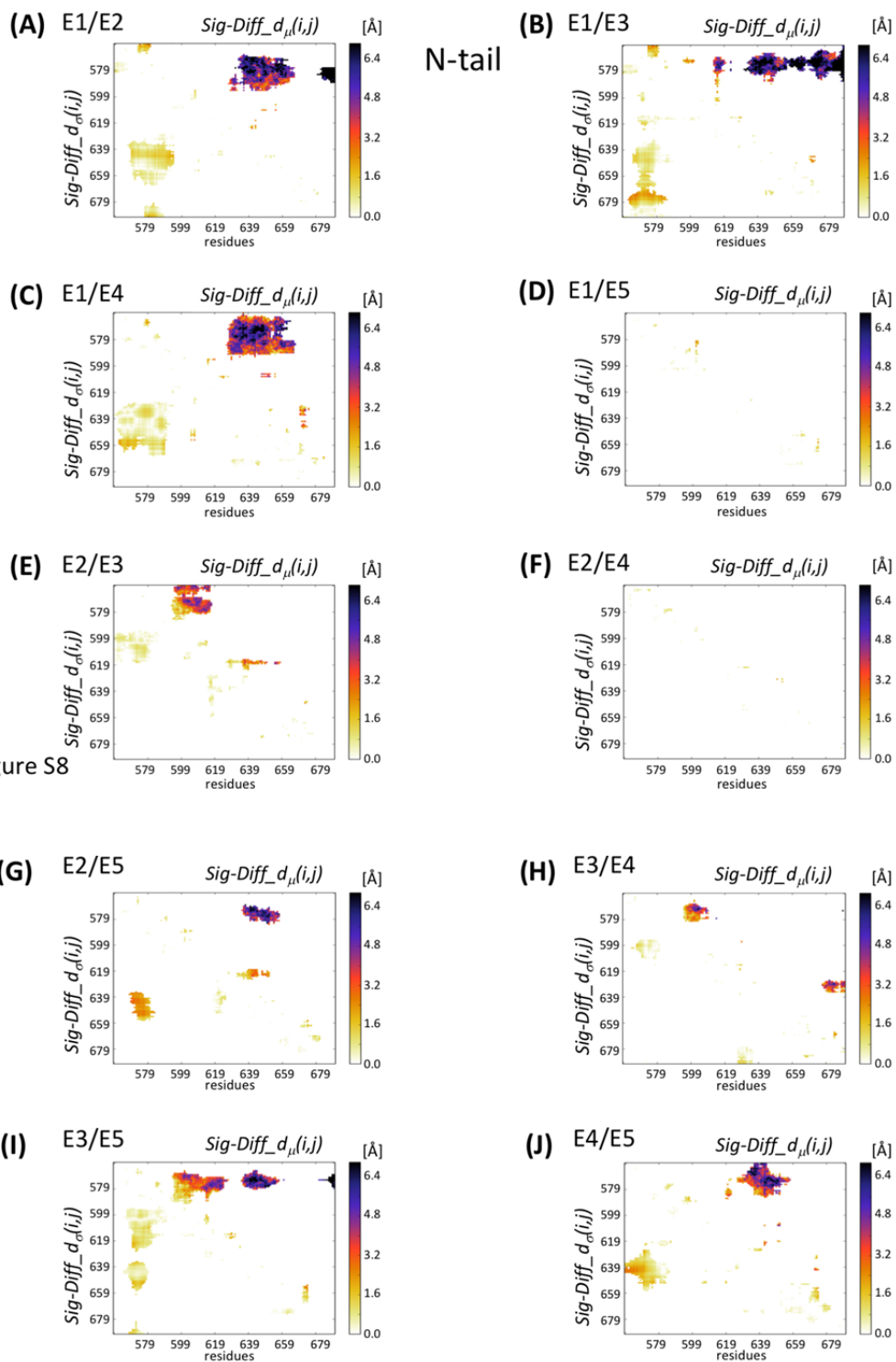


Figure S8

Figure S8

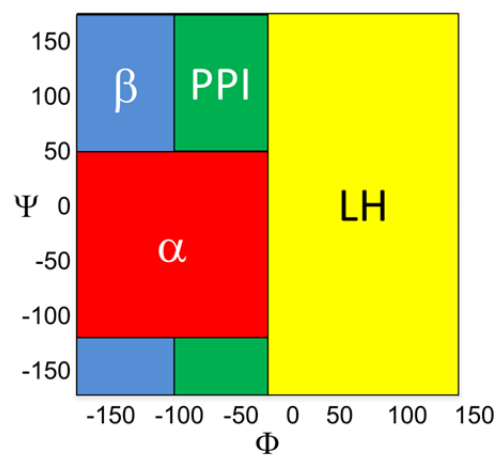


Figure S10